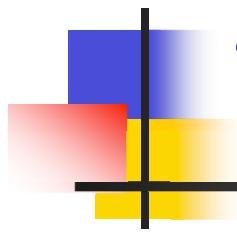
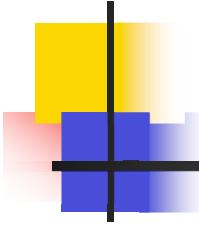


Proposal for dCache based Analysis Disk Pool for CDF

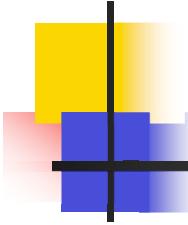


presented by
Doug Benjamin
Duke University
on behalf of the CDF Offline Group



Outline

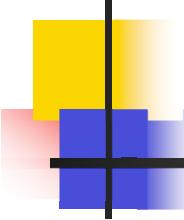
- The Problem & Proposed Solution
- Deployment Process
 - Collaborative and phased approach with feedback from users and experts at each stage



How CDF does data analysis

Collision Data Flow

- Raw data written to tape
- Measure the final calibrations
- Full and final reconstruction on production farm or CDF user analysis cluster (CAF or FermiGrid)
 - Input read in from tape and output written to tape
- Centralized ntuple production (most CDF physicists access the data through ntuples)
 - Collaboration wide ntuples produced on production farm or CDF CAF
- Physics groups further reduce the centralized ntuples or production files (strip off types of events) for further study
- User analysis of ntuples and production data run on CAF or FermiGrid
- Final Analysis on user desktops or institutional machines



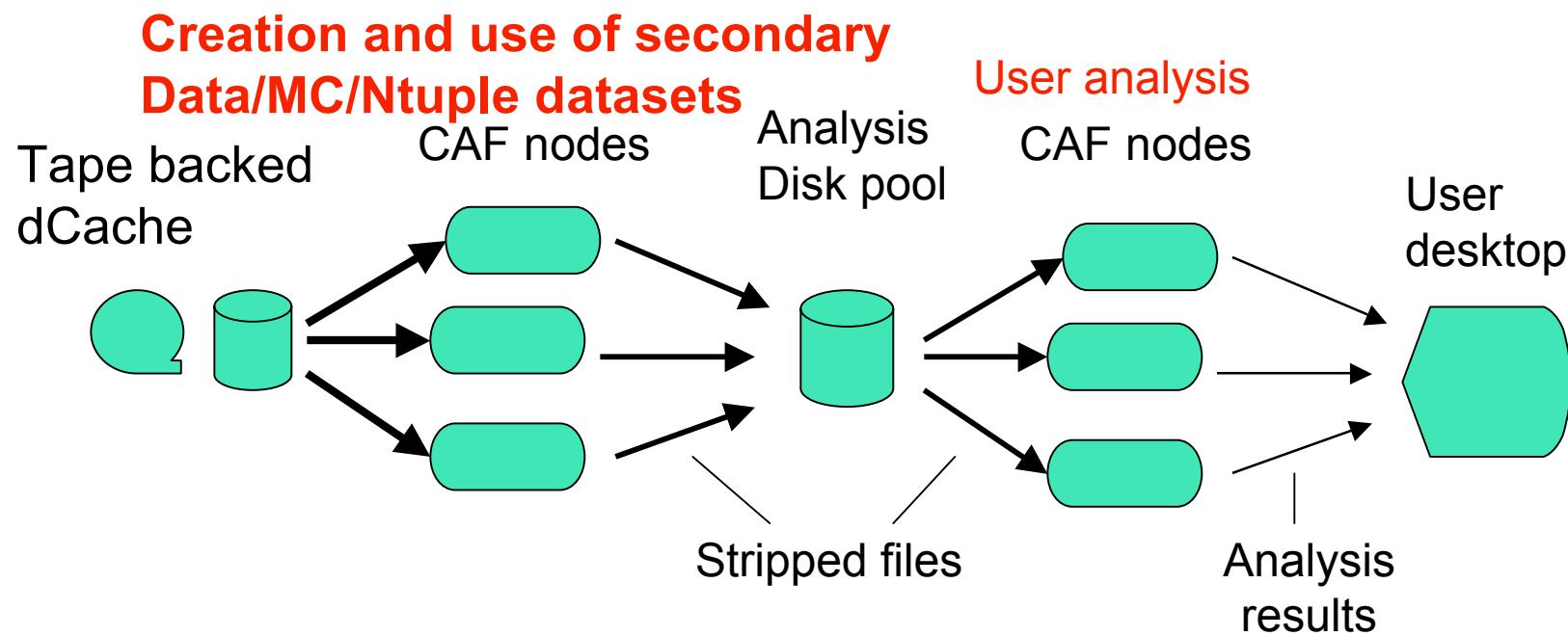
How CDF does data analysis (2)

Simulation Data Flow (Monte Carlo)

- Monte Carlo generation and reconstruction on remote computing
 - Offsite dCaf's
 - NAMCAF (OSG sites - including Fermi Lab)
 - LCGCAF (Europe)
- Files return to FNAL for storage on archival media (tape)
 - Small files (< 1 GB in size) must be merged together
- MC files written to tape
- Centralized ntuple production run on MC Samples
 - Some physics groups produce ntuples at time of generation/reconstruction
 - Other groups create ntuples after the MC samples are on tape
- MC ntuple files are merged together before being written to tape
- Like Collision data - Physics groups further reduce the centralized ntuples or \MC files (strip off types of events) for further study
- User analysis of MC ntuples and data performed on CAF or FermiGrid
- Final Analysis on user desktops or institutional machines

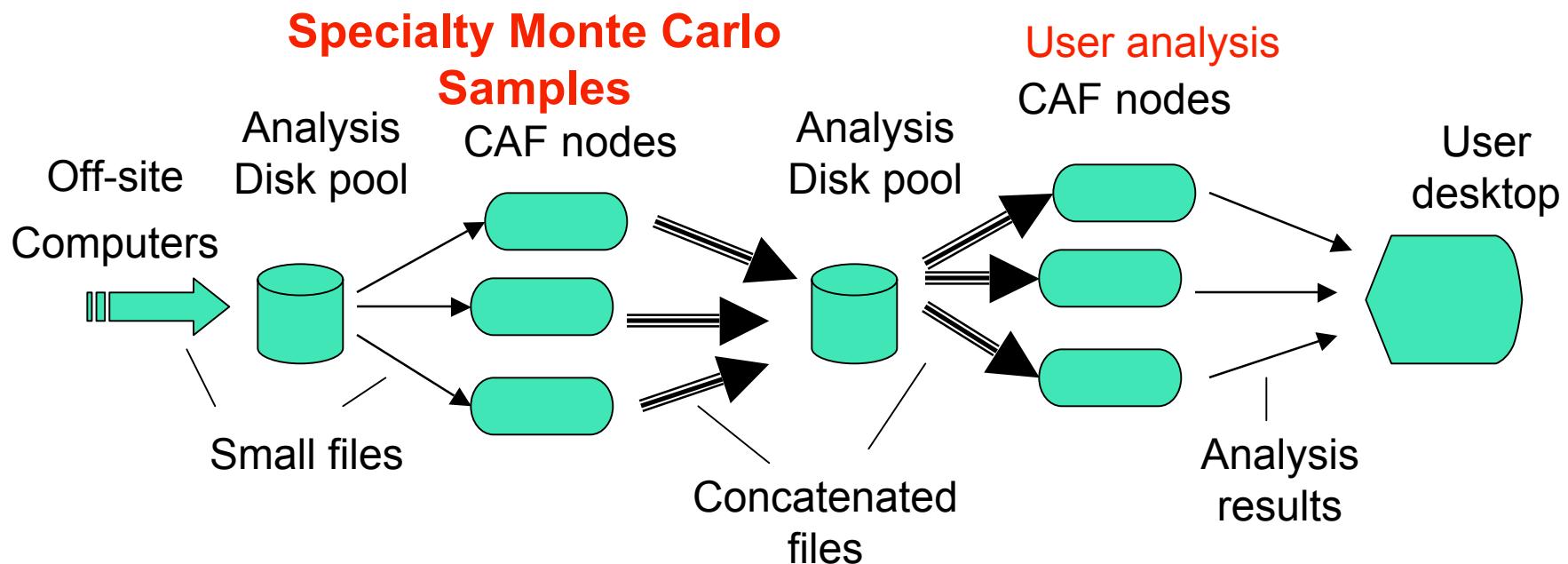
Physics groups Project Disk space needs

- Physics groups need project space
 - 5 Groups: B , Top, Exotic, QCD, EWK
- Several use cases :
 - secondary or tertiary datasets creation and use



Small Analysis Group disk space needs

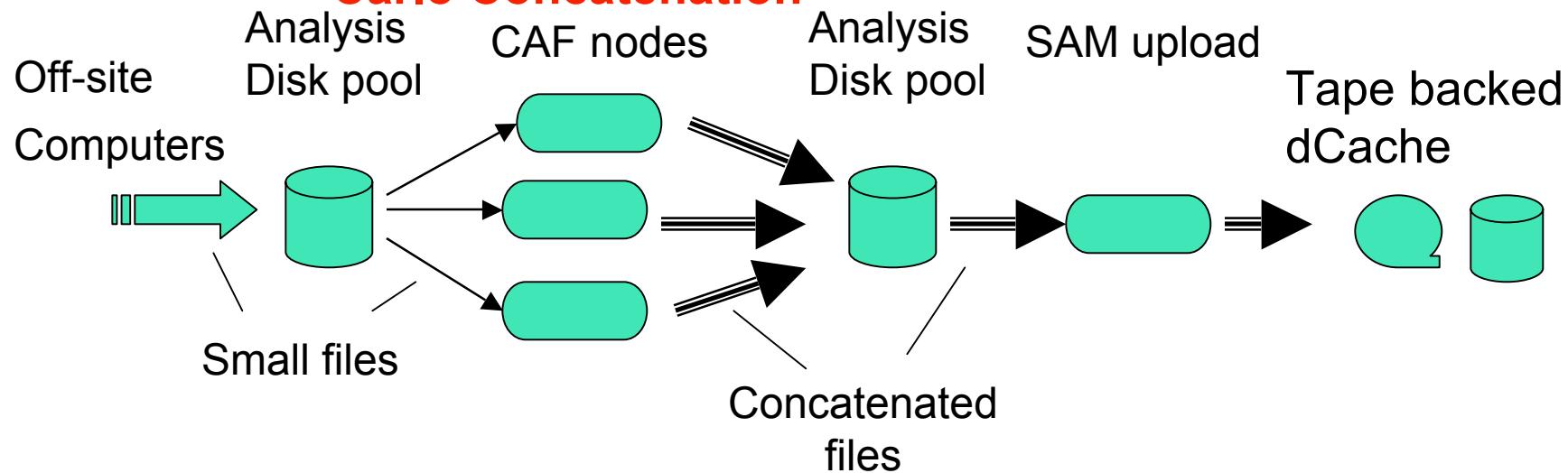
- Small Analysis Group project space
 - (a few people working on an specific analysis)
 - Analysis specific MC - Special study samples

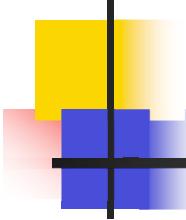


Offline Monte Carlo Group disk space needs

- Concatenation of MC data files and ntuple files produced offsite
 - File size requirements in data handling system (> 1 GB)
 - Maximum CPU time requirements of Monte Carlo jobs limit output file size
 - Implies files must be concatenated before uploading into DH system

Production Monte Carlo Concatenation

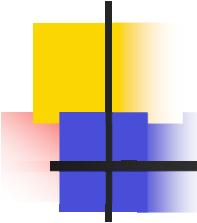




Data to be stored

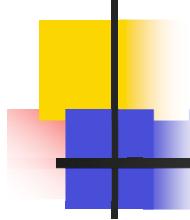
- Secondary and Tertiary datasets
 - production data, MC data and ntuples
 - Stripped from data held in data handling system
- Temporary data
 - Data used during the concatenation process
 - MC data used for small analyses that can be easily regenerated

(Note - archival quality data will be stored in the data handling system)



Current situation

- CDF is using discrete independent fileservers.
 - ~70 fileservers for project space (> 160 TB)
 - Can not easily Centrally manage the space
 - Space inefficiently used (stale data)
 - Data migration requires access to individual nodes and users have to change all their scripts
 - Individual files servers approach not scalable
 - Users often overload one or two file servers
 - No Load balancing



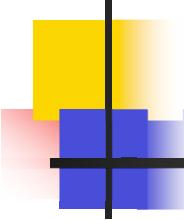
Physics groups space needs

- Polled physics groups to get space needs per fb^{-1}

groups	space per fb^{-1}
Top	15 TB
QCD	12 TB
Exotics	20 TB
B group	8 TB
Electroweak	10 TB
MC production	20 TB
Total	85 TB/fb^{-1}

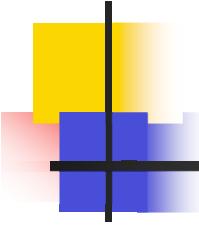
160 TB is sufficient for data taken to date

400 - 600 TB should be sufficient through FY '09



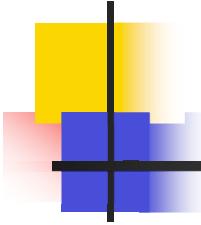
Project Disk requirements

- Stability, simplicity and reduced operational effort
- Scaleable - (add space/users as needed)
- Centralized administration to monitor and maximize our limited resources
- Simple user code changes
 - (avoid lose in productivity)
- Decouple file storage from compute nodes
 - Network file access (URL based)
 - Avoids strong coupling between CAF worker nodes and fileservers hosting data



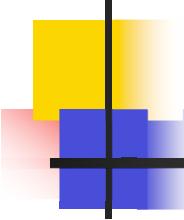
Proposed Solution

- Replace the majority of the static disk space with dCache based pool (the analysis disk pool)
 - Use it for large files for which dCache is known to work well
 - Store small files, e.g. log files on other disk based system e.g. on nfs mounted disks visible from Interactive Login Pool nodes
- dCache analysis diskpool finished prototype testing phase
 - 4 head nodes for the services
 - 9 fileservers (~ 50 TB total)
 - 5 oldest filesersevers (~ 4 years old) to be replaced after this review



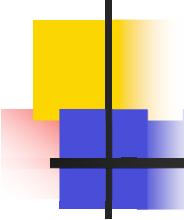
Advantages

- Solution adopted/supported at Fermilab and within HEP
 - Allows for unified support and expertise sharing
 - groups using dCache (DESY, CMS, CDF...)
- Global space management
 - more efficient use of disk resources
 - more transparent maintainability
- Decoupling of the name and disk space
- Scalability
 - User files spread across several servers
- Client software already used by CDF



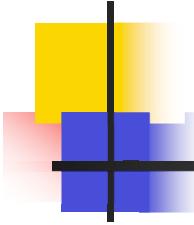
Risks and Risk remediation

- In a centralized system a small group of users may inadvertently affect all users
 - User education reduces the risk
 - Have not had any user induced outages to date
- Added software layer on top of storage hardware
- Centralized resources stability
 - Production dCache system fairly stable (some problems with fileservers - solved by proper hardware drivers and retirement of the oldest nodes)
- File server stability can affect large number of users.
 - Spread data across several servers
 - Minimizes data inaccessibility due to one server down
 - Balances the load



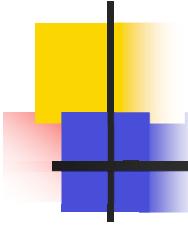
Risks and Risk remediation(2)

- What if some component of dCache changes underneath us (like namespace change: PNFS to Chimera)?
 - Not a problem Chimera has a filesystem core onto of a relational database. To the CDF end user he/she won't see much of a difference
- What if dCache is replaced with technology Y?
 - If a root TFile class for technology Y exists then minor change to CDF software (URL's to access files across the network)
 - If namespace looks like a filesystem then minimal change to CDF software



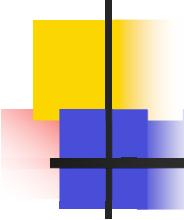
Risk Management

- Many fileservers to reduce the dependence on only a few parts.
- All data stored on system could be restored or regenerated with reasonable amount of effort.
- Monitor the system to make sure it stays within the stable limits
- Establish usage guidelines and limit exposures when possible (e.g. limit pnfs mounts)
- Use good quality hardware for head nodes (have a spare node capable of being put into service quickly)



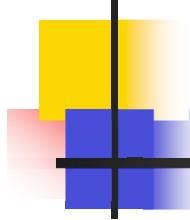
Usage guidelines

- Minimum file size 1 GB for long lived files
 - Small files to be concatenated are excluded
- Prohibit use of tar, unzip and similar operations on Disk Pool
- Prohibit compilation, linking, debugging and similar activities
- Restrict large scale use of ls on pnfs namespace as a general user option.
 - Browsable listings of directories and files will be created at least once per day and visible on the web



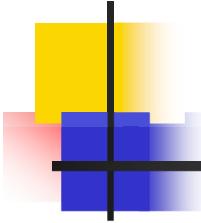
Usage guidelines(2)

- Prohibit the storage of log files in Disk Pool
 - Log files are to be stored on static disks in the CDF interactive login pool (ILP)
- Prohibit looping over multiple files at an unduly high rate (for example TChain in Root or other techniques)
 - File transaction rate should be limited to 5 Hz across the entire system
- Users agree to use the Disk Pool within the established technical specifications



Production system Specifications

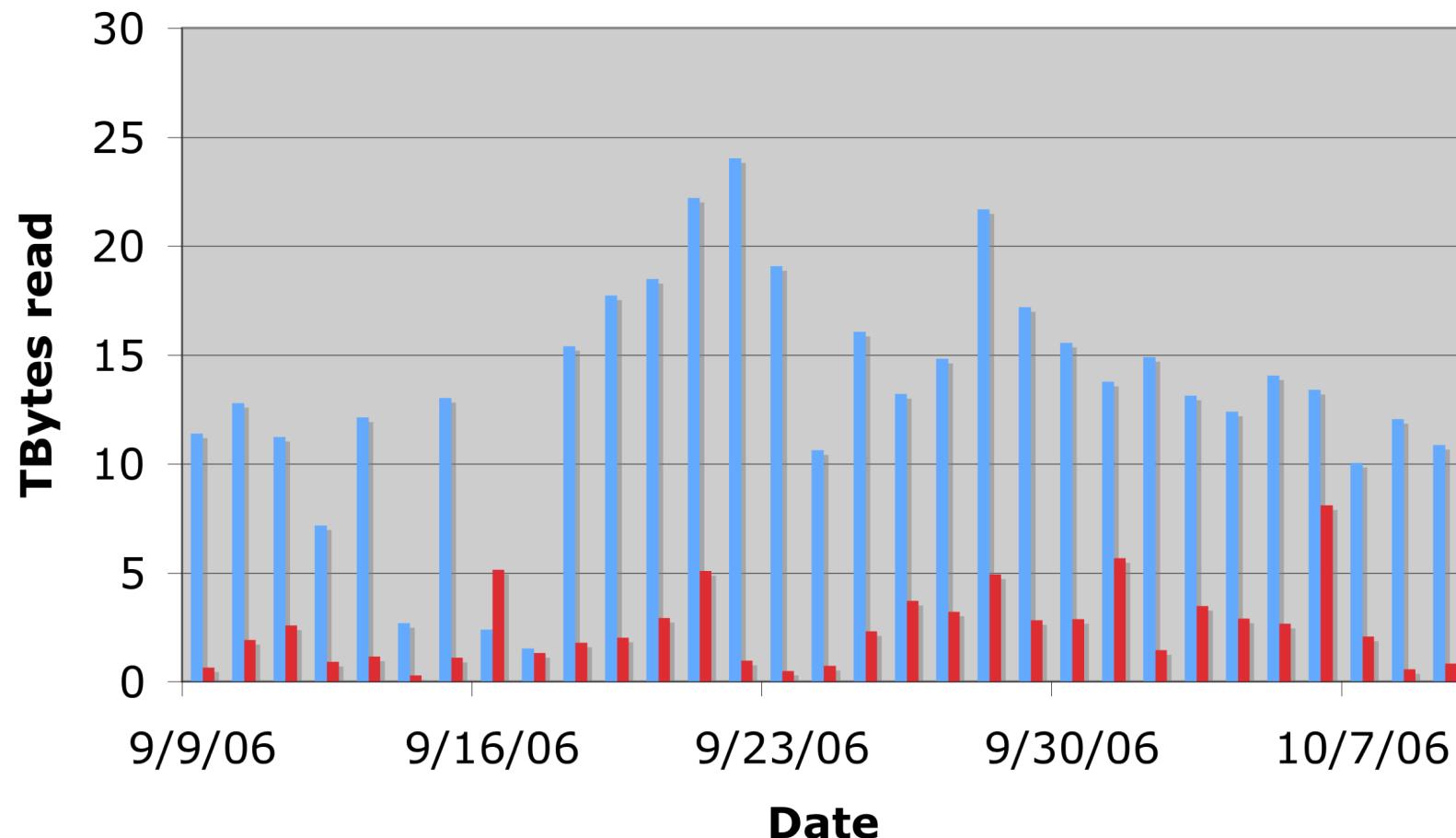
- Total data volume served per day - 17 TB/day
 - Based on average current University File server system (150 MB/s) and prototype disk pool system (40 MB/s)
- Aggregate bandwidth - 250 MB/s
 - Peak of University system - 170 MB/s and prototype disk pool - 80 MB/s
- Maximum supported transaction rate 5 Hz
 - Expected transaction rate ~ 0.2 Hz
- Maximum number of concurrent jobs 700



Daily data volume (TB) by CDF dCache systems

Daily Amount of data read

■ Production dCache ■ Analysis Disk Pool



Analysis Disk pool - # of CDF jobs/hour

Total Active movers
(jobs) - blue line

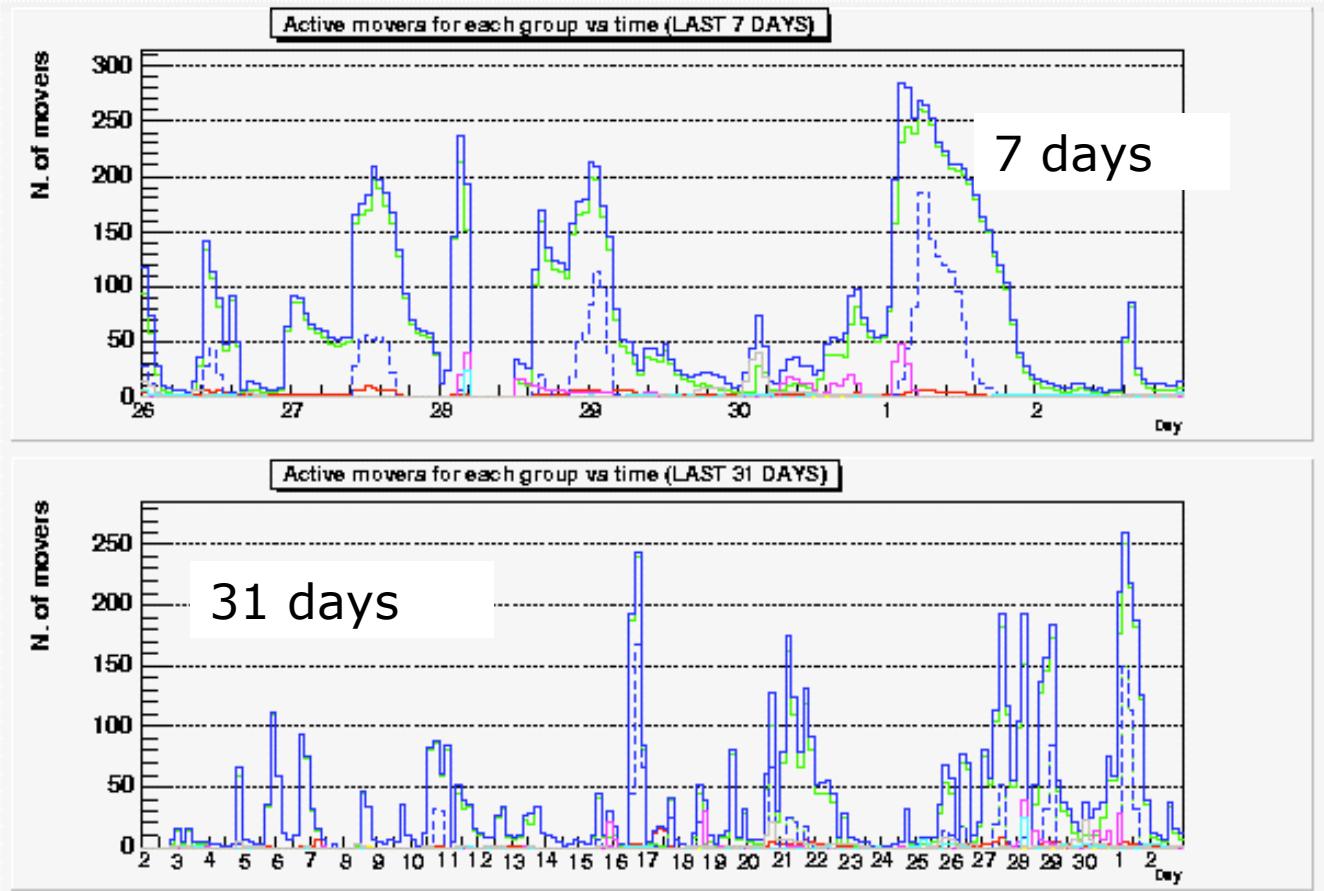
< 300 jobs per hour

Total data volume

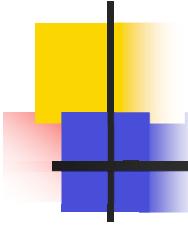
~ 50 TB

Peak read rate:

< 120 MB/s

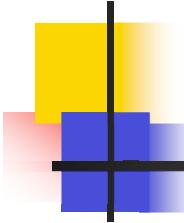


Note - number of jobs/ transaction rate will scale with amount of CPU cycles available (worker nodes) - (not data volume stored) - Expected CDF CPU growth 200% now to FY'09



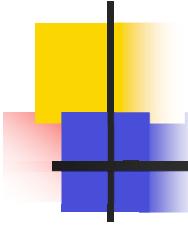
Proposed Support Model

- Three tier approach:
 - Day-to-day operations and trouble shooting by CDF power users & CDF offline operations managers
 - Diagnosing difficult problems, evolving and reconfiguring system when needed by a group from a CDF institution (per a to-be established MOU) – also serving as a point of contact between the experiment and CD
 - Expert level consultations within to-be agreed upon limits by CD dCache development team



Proposed Support Model (cont'd)

- Hardware, OS support by REX/System Administrators
 - Head nodes same level of hardware support as tape backed dCache system
 - File servers - support 8x5 - similar to all other file servers used by CDF
- CDF experiment provides primary support for the dCache software



Personel Requirements (in FTE)

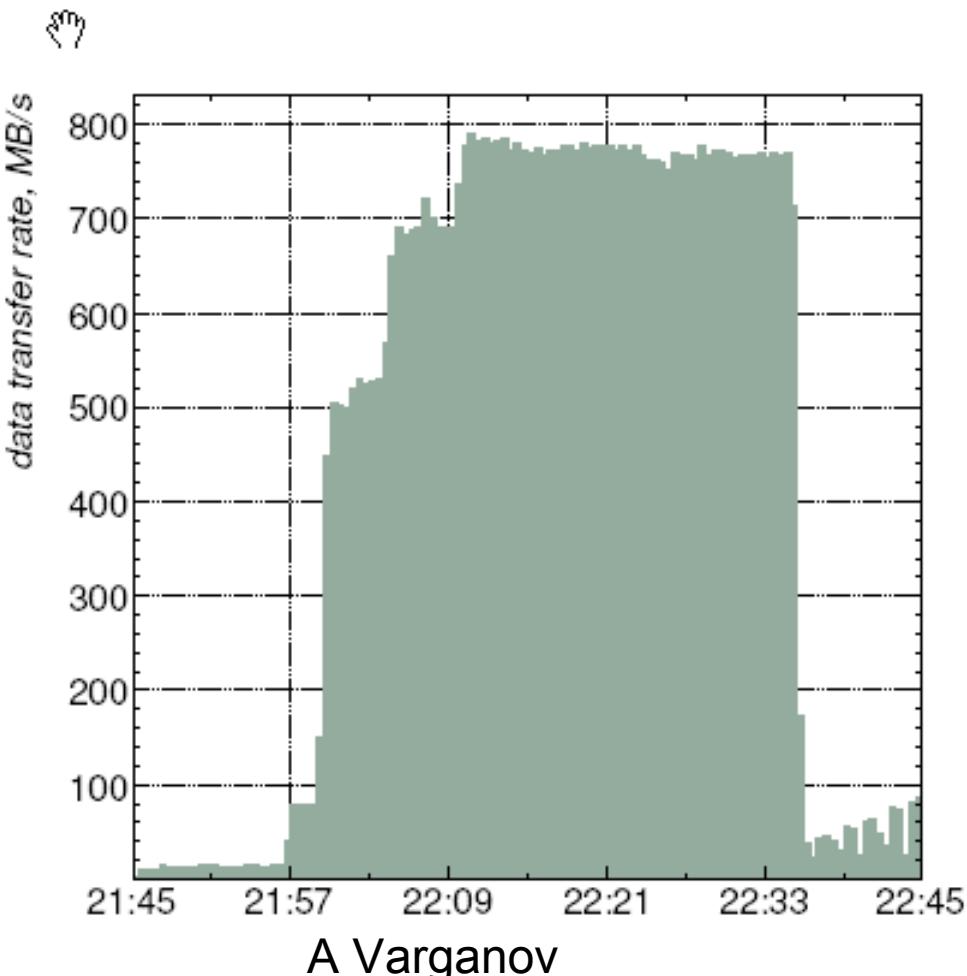
- CDF

- < 0.25 FTE primary support
(Alexei Varganov will provide examples in next talk including his work load)

- Computing Division

- REX system administrators
 - Twice the effort spent on production dCache head nodes (We have 2 systems now) (*Small amount of effort*)
 - dCache developers
 - < 10 hours per month for consultation

System testing

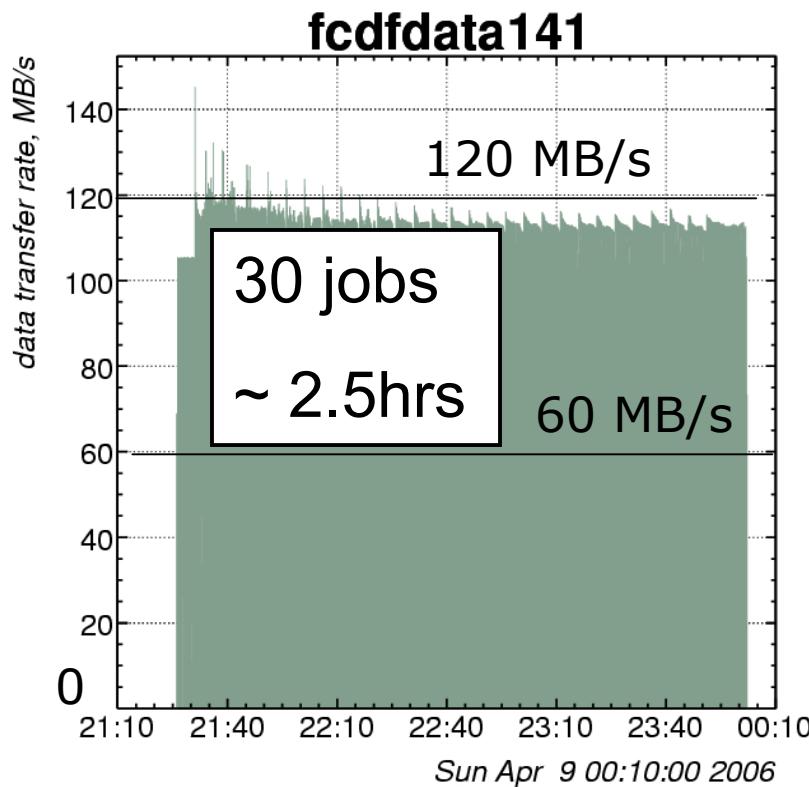


Fileserver read performance

- Aggregate read rate
~ 800 MB/s entire system
- ~ 30 minute test
- Network limited
- Dccp used to move files

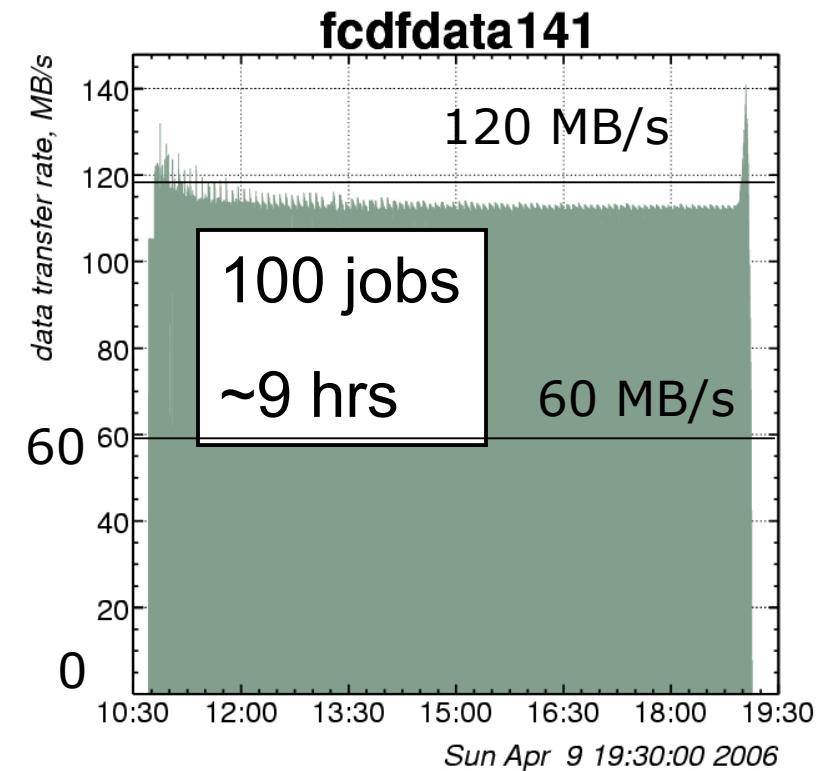
A Varganov

Read test - Job scaling



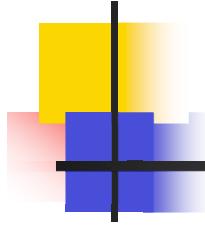
Read Transfer rate (MB/s)

Network limited (both tests)

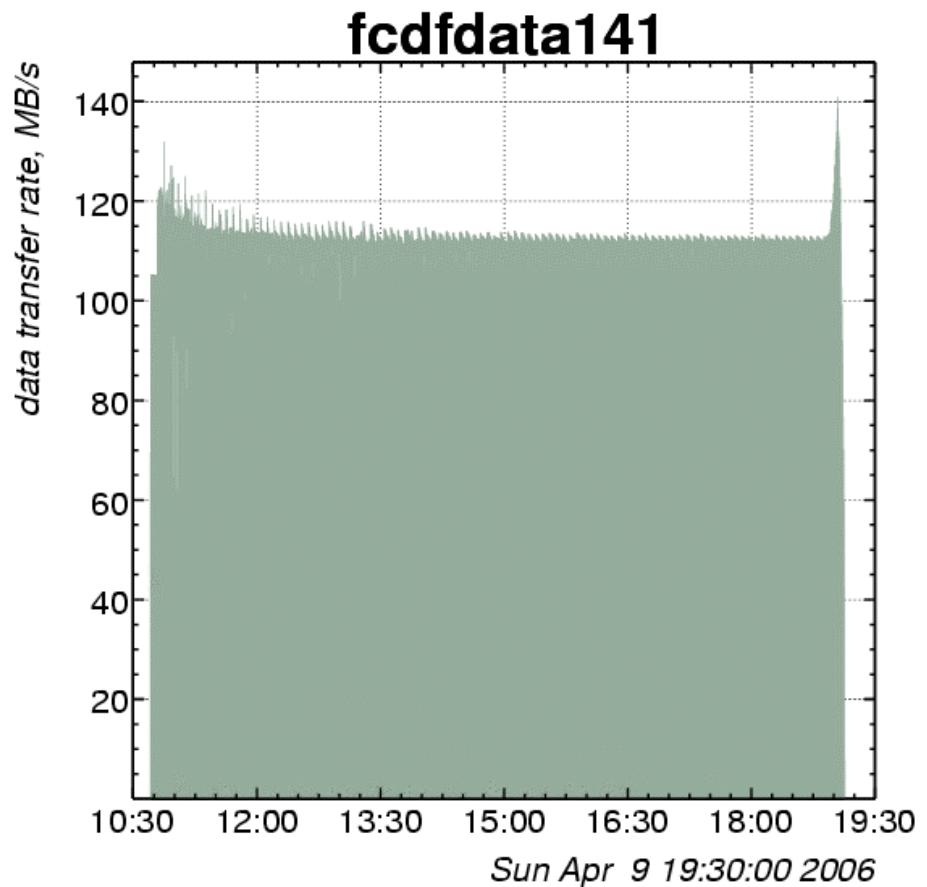
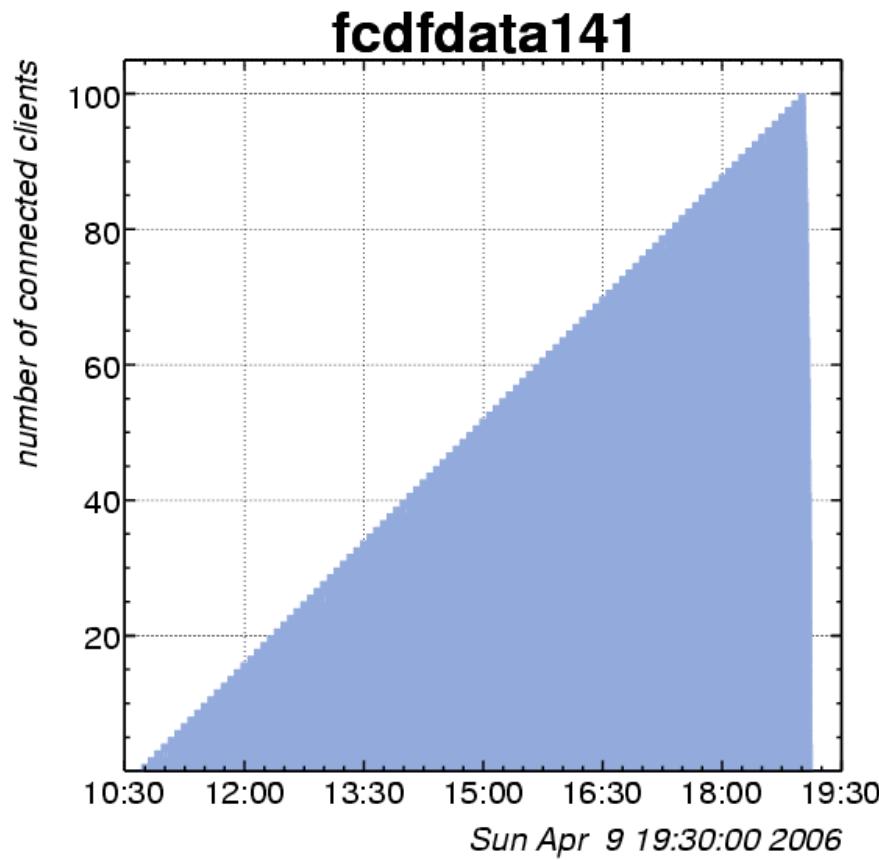


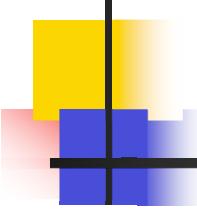
A Varganov

Nexsan ATA beast file server

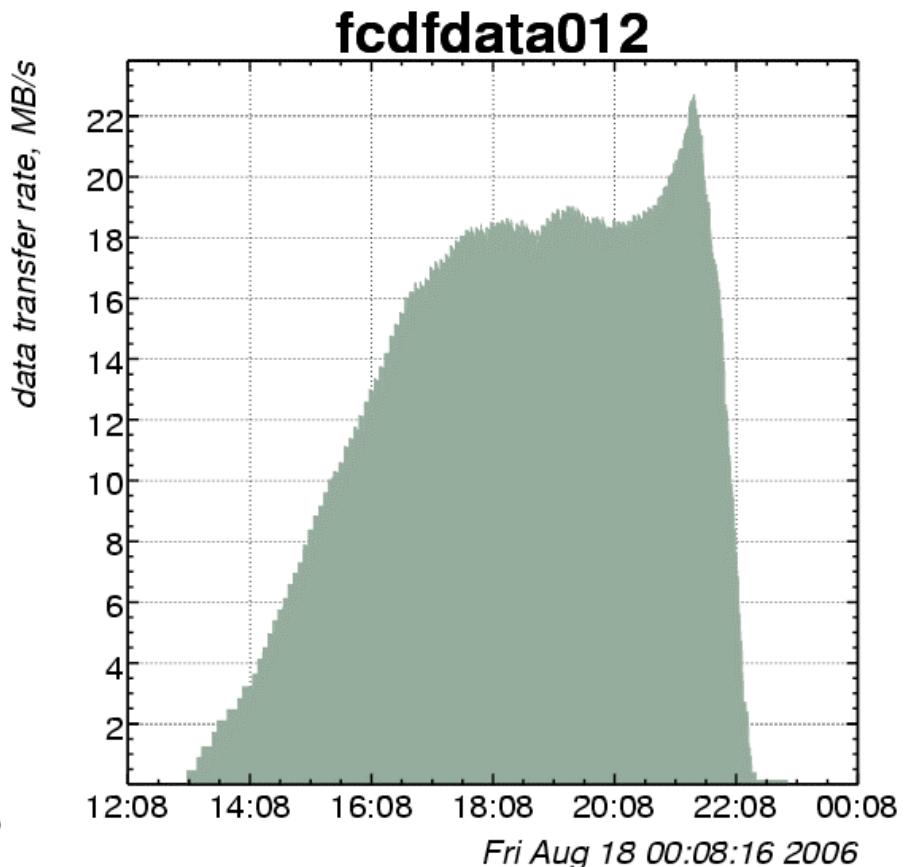
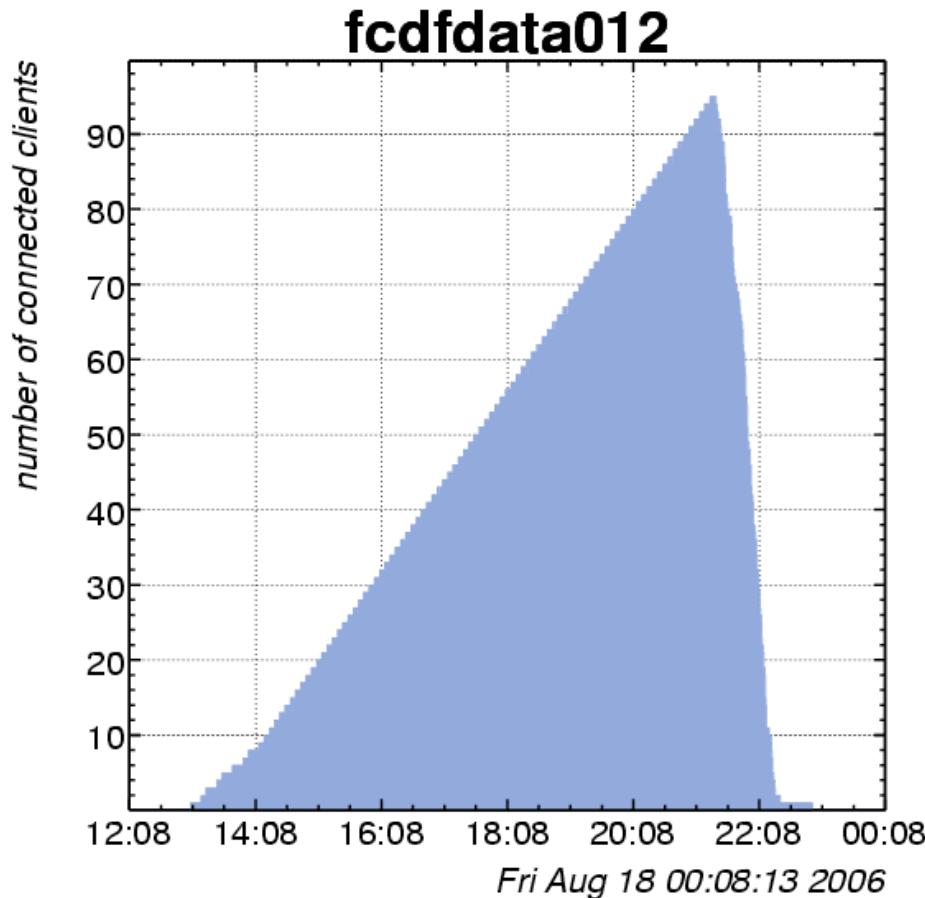


100 job read performance test (dccp)





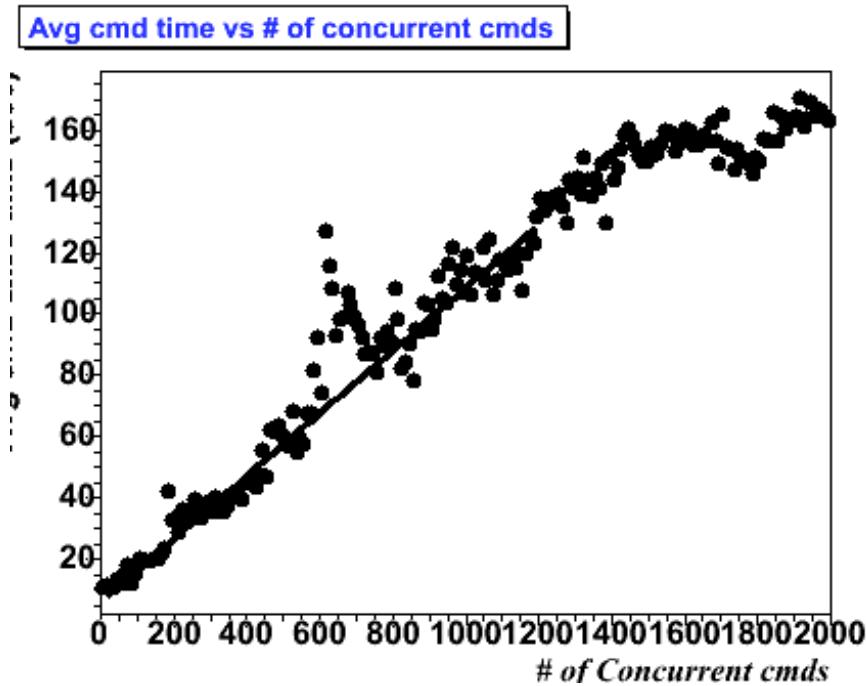
Root file read performance



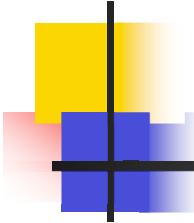
Note: Root files are compressed on disk - (decompression takes time)

Old fileserver used in test

PNFS bandwidth testing



- Used both dCache prestage command and root Tfile open to make measurements.
- Found response linear up to ~ 1500 simultaneous clients ~ 0.1 secs per Tfile open
- **Automatically recovered** when test finished (had > 2000 simultaneous clients)



Summary

- In order to produce excellent physics results - CDF needs project disks space
- dCache provides an excellent choice for CDF
 - CDF has extensive experience with dCache
 - Minimizes change to users code - reduce impact on users to increase their efficiency
 - Minimizes the operational effort